D516–D519 *Nucleic Acids Research, 2004, Vol. 32, Database issue*
DOI: 10.1093/nar/gkh111

HGVbase: a curated resource describing human DNA variation and phenotype relationships

D. Fredman, G. Munns, D. Rios, F. Sjöholm, M. Siegfried, B. Lenhard, H. Lehväslaiho[1] and A. J. Brookes*

Center for Genomics and Bioinformatics, Karolinska Institute, Berzelius väg 35, S-171 77 Stockholm, Sweden and [1]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received September 15, 2003; Revised and Accepted October 14, 2003

ABSTRACT

The Human Genome Variation Database (HGVbase; http://hgvbase.cgb.ki.se) has provided a curated summary of human DNA variation for more than 5 years, thus facilitating research into DNA sequence variation and human phenotypes. The database has undergone many changes and improvements to accommodate increasing volumes and new types of data. The focus of HGVbase has recently shifted towards information on haplotypes and phenotypes, relationships between phenotypes and DNA variation, and collaborative efforts to provide a global resource for genome–phenome data. Open sharing and precise phenotype definitions are necessary to advance the current understanding of common diseases that are typified by complex aetiologies, small genetic effect sizes and multiple confounding factors that obscure positive study results. Association data will increasingly be collected as part of this new project thrust. This report describes the evolving features of HGVbase, and covers in detail the technological choices we have made to enable efficient storage and data mining of increasingly large and complex data sets.

INTRODUCTION

As human genome disease studies increase in both scale and scope, there is an ever-increasing need for effective informatics platforms that summarize all aspects of human genome variation. Our database project has been maturing over many years in an attempt to meet that need. The project name, 'HGBASE' (Human Genic Bi-Allelic SEquences), was changed in 2001 to 'HGVbase' (Human Genome Variation database) to reflect a deeper relationship with the Human Genome Variation Society (HGVS) and our focus upon the whole genome rather than just genes. Currently, the database home pages (http://hgvbase.cgb.ki.se) are accessed by >2000 computer domains every month, and perhaps by twice as many individual scientists.

Presently, HGVbase attempts to capture, filter and enhance all described human genome sequence variation in the public domain. This includes single-base and multiple-base differences, whether suggested or proven, regardless of functional consequence or existence of frequency data in any population. Records are assigned a unique and permanent HGVbase identifier to facilitate database cross-referencing and publication. Data collection efforts are broad and not restricted to submitted data. We actively seek out a range of other data sources that would otherwise be lost to the bioinformatics future. Similarly, not all submitted data is accepted, since our curation processes apply stringent quality criteria to the information we receive and/or gather. On average ~30% of public-domain single nucleotide polymorphisms (SNPs) are not represented in our project because of inconsistent or very low-quality data, which results in HGVbase being enriched for SNPs more likely to be real (i.e. polymorphic) in tested populations.

RECENT DEVELOPMENTS AND DATABASE CONTENT

Genetic variation

HGVbase continues to gather information pertaining to thousands of SNPs every year from the literature, our own and collaborative discovery efforts and unsolicited submissions. These SNPs are typically in genes potentially related to common disease, and they often come with population frequency information. We are also keen to gather and receive 'negative findings' where predicted SNPs were tested in one or more populations but not found to be polymorphic—valuable information that often fails to be reported elsewhere. Database exchange of core information with dbSNP (1) ensures that we incorporate data from high-throughput discovery efforts. HGVbase release 15 contains information on 3 million SNPs, of which 29 000 are found in 10 000 genes and 41 000 have allele frequency information.

Gene structure

The location of each represented variant is presented in the context of available gene predictions, and SNPs within or around genes are described as 'exonic', 'intronic', 'utr' or 'flank' (within 2 kb of the gene boundary). HGVbase currently considers only genes with a HUGO nomenclature committee approved definition (2), as represented in the Ensembl database (3). Non-synonymous SNPs are grouped into three

*To whom correspondence should be addressed. Tel: +46 8 7286630; Fax: +46 8 324826; Email: anthony.brookes@cgb.ki.se

broad classes based on their predicted effect on the protein level: 'benign', 'possibly damaging' and 'probably damaging'. The methods used for these functional predictions are described in (4,5). We are currently investigating the possibility of also including information on SNP relationships with transcription factor binding sites.

Haplotype data

In contrast to other SNP databases, alleles as well as SNPs in HGVbase carry unique identifiers. This allows haplotypes to be easily represented as lists of allele IDs, avoiding the inevitable phasing problems that arise when representing haplotypes as lists of markers with assigned strand parameters and base types. Similar to SNPs, haplotypes are associated with frequencies, cross-referenced and used in genotype definitions. With large discovery efforts underway that promise to deliver genome-wide coverage of common haplotypes (6) we expect haplotypes and genotypes to become the principal functional units for the study of correlations between sequence variations and phenotypes. In release 15, HGVbase holds 15 000 haplotypes on six different chromosomes.

Phenotype data

In anticipation of sophisticated data standards (see Future development) we have created a prototype system for handling phenotype data in HGVbase. Database structures and interfaces for input and curation of data revolve around curated controlled vocabularies, with limited hierarchical structure. This allows us to handle phenotype descriptions and their relation to human polymorphism, submitted or gathered from LSDB databases and clinical diagnostics laboratories. Harvesting such data is part of a planned joint venture with the HGVS. We are also developing a user-friendly Windows program written in Visual Basic that researchers may use to help submit genotype, phenotype and genotype–phenotype relationship data.

Low-quality data: duplicons, paralogues and repeat elements

Of all data passing our acceptance criteria for inclusion in HGVbase (7), 11% are such that they should generally not be used in genotyping experiments. These SNPs are in repetitive sequences that are impossible to assay in a straightforward manner. We identify them by comparing their mapping positions on the chromosome with those of duplicated segments (8), by scanning for common repeats (9) and by noting SNPs whose allele frequencies are 50/50 in all reported populations. We flag such variation as 'low quality' to give users the possibility of excluding such variations from their laboratory studies. Haplotypes that contain these low-quality SNPs are included in HGVbase, though their validity must be questionable.

Genomic locations

We now use NCBI-produced coordinates for the corresponding reference SNP clusters as the genomic locations for HGVbase SNPs. Whilst some studies have suggested that the NCBI SNP mapping information is sometimes misleading (10), we believe that the benefits of keeping a consensus between SNP databases on the location for each SNP avoids confusion in an already problematic field. Nevertheless, we do flag questionable assignments, and provide feedback to the NCBI on suspect mapping information. Mapping information is re-considered in every new release of HGVbase.

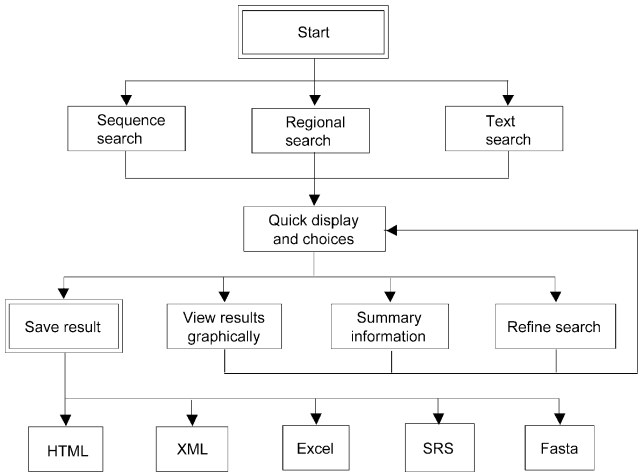

Figure 1. HGVSearch allows iterative search for entries in HGVbase. Each box represents a search stage with a corresponding dynamically rendered web page, and the arrows display how the user may proceed through the search procedure. Boxes with double walls are start and end states. The system allows data display in convenient formats for intermediate inspection or saving.

QUERYING THE DATABASE

HGVbase online searches have until now been performed through various disparate systems and interfaces. Search by sequence was available through Fasta3 (11) on the EBI site. The SRS tool there (12) has served as a powerful text-based retrieval system but it fails to deliver functionality needed to search and display genetic features with reference to chromosomal locations. To facilitate positional searches we previously developed a simple wrap-around SRS, but due to technological obstacles it was limited to the retrieval of 250 HGVbase entries at a time.

To counter these limitations and provide a homogeneous environment for all queries against HGVbase, we recently developed a new retrieval system (HGVSearch) that is accessible through web interfaces and command-line tools. For maximum flexibility, we decided on an iterative search procedure allowing for continuous visual inspection and search refinement, and included the possibility of viewing and outputing results in a variety of convenient formats. The iterative environment is described in Figure 1, and has three major entry points: (i) keywords or batches of SNP and/or haplotype identifiers, (ii) chromosomal position defined by exact location, chromosomal band or genetic marker, and (iii) one or more sequences. Given any of the above inputs the user is presented with the number of HGVbase items found and the option of viewing, refining or saving the result. Intermittent results can be viewed either graphically or in a summary table. Refinement can be used to limit the data set to entries with certain properties, such as allele frequency, genic location or quality status as shown in a snapshot of the HGVSearch web interface in Figure 2. Results can be saved in XML, HTML,

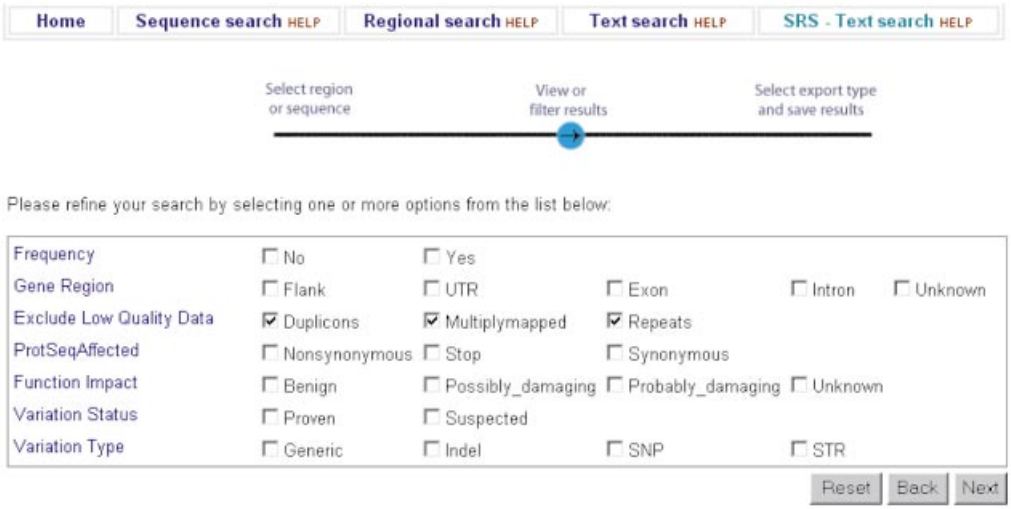

Figure 2. The web interface for HGVSearch shows the different entry points for a search in the top list and a time-line below tracking progress through the search procedure. The current stage, 'Refine search', here shows the available options by which a search can be limited. The options to exclude low-quality data are pre-selected.

SRS, flat file and Fasta formats. HGVSearch can also be used for ID-based batch retrieval of many entries at once. There is no volume restriction on downloads, although structured representations (e.g. XML) of whole-chromosome data may take several minutes to render. Researchers who wish to acquire the whole of HGVbase for local use with the HGVSearch API (or via incorporation into their own system) can download MySQL tables from our FTP server (ftp://ftp.ebi.ac.uk/pub/databases/variantdbs/hgvbase/).

SYSTEM DESIGN AND IMPLEMENTATION

From the production copy of HGVbase we have produced a search-optimized relational database that serves as the HGVSearch back-end. The HGVSearch database contains a mix of positional tables holding genomic start–stop coordinates for HGVbase and relevant Ensembl features, non-positional tables with data independent of position and aggregate tables used for optimized data retrieval. Minimal data types are used to represent the information, and indexes are optimized to fit in the primary memory of the server whenever possible.

We considered remodelling the HGVSearch database into a warehousing star-schema model (13). Due to the high cardinality of certain properties (e.g. when frequencies from many populations have been submitted), this solution was only feasible if we resorted to representing summary information, which is not compliant with the objective of presenting all pertinent data for each entry in HGVbase.

The HGVSearch web application consists of a Perl API communicating with the database back-end, and Perl cgi scripts that serve HTML pages and render graphics using the Bio::Graphics library (14) through an Apache server running on Redhat Linux 9.0. Keyword and positional searches are performed against the MySQL database, while sequence searches utilize the BLAT package (15) for quick retrieval of near-exact matches in the human genome. For each iterative search session, state is maintained through cookies and hidden parameters, and the current result set is stored in a temporary table held in the primary memory on the MySQL server. Keeping the current result set on the MySQL server is inexpensive and limits the level of data transferred between the client and the server. It also optimizes performance when refining data sets because the transient result table is small and able to rapidly create new joins with other tables.

The central data export format for HGVbase is XML, from which SRS, Text, HTML and MS Excel representations are created through transformation of the XML structure. We also use XML for data exchange with other databases (e.g. dbSNP). Having several years of experience with XML for genetic variation applications, we are now actively supporting OMG efforts to create standards for the exchange of DNA variation data in the SNP community. Should the need arise, we could also use this XML to offer SNP data sharing via open standards such as the Distributed Annotation System (16) or BioMOBY (17).

FUTURE DEVELOPMENTS

Whilst we and the community have mastered the 'uni-dimensional' challenge of characterizing DNA sequence (including causal disease mutations), a greater challenge lies ahead of us: shaping effective exploration into genome function and how this relates to any and all observed phenotypes. Our focus in the coming years will be upon improving and extending the ways in which HGVbase stores and represents phenotype data, and its dependence upon genome sequence differences. Our platform of database search functions, submission tools and curation assistance software, is now being extended to fully incorporate sophisticated phenotype information content. Important input will be drawn from the more than 40 genomics and bioinformatics groups interested in database representation of mutation and phenotype data assembled under the project title 'PhenoFocus' (http://www.phenofocus.net). Included are many groups working on model organism databases, and the phenotype

challenges therein. Via web and email modes, PhenoFocus will create and disseminate broad consensus solutions for computational handling of phenotype data.

Other planned activities will revolve around leveraging information on haplotypes and continuing to provide value added annotation such as functional predictions and genotyping assays on a variety of platforms.

AVAILABILITY

All HGVbase data are freely available for academic research purposes, and data and accompanying descriptions and software are available for download from our website (http://hgvbase.cgb.ki.se). We make no claim of ownership or warranty for supplied data, though our specific compilation and representation of it are subject to clear copyright and usage principles (ftp://ftp.ebi.ac.uk/pub/databases/variantdbs/hgvbase/LICENSE). These policies are designed to ensure that the HGVbase resource remains freely available to everyone for research purposes.

ACKNOWLEDGEMENTS

We acknowledge Erik Lönroth and Gergely Hajdu for work on the HGVSearch web interface, Shamil Sunyaev and Pauline Ng for functional predictions and users of HGVbase for valuable submissions and suggestions. We thank Thermo-Hybaid (Interactiva division, Germany) for support during early database development and for transferring the database to the public domain. We also wish to thank the KI, EBI, EMBL, Pfizer and Celera for their unconditional financial and practical support of the HGVbase project. D.F. is supported by a grant from KK-Stiftelsen through the Research School of Medical Bioinformatics.

REFERENCES

1. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
2. Wain,H.M., Bruford,E.A., Lovering,R.C., Lush,M.J., Wright,M.W. and Povey,S. (2002) Guidelines for human gene nomenclature. *Genomics*, **79**, 464–470.
3. Hubbard,T., Barker,D., Birney,E., Cameron,G.,Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.
4. Ng,P.C. and Henikoff,S. (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.*, **31**, 3812–3814.
5. Ramensky,V., Bork,P. and Sunyaev,S. (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.*, **30**, 3894–3900.
6. Gabriel,S.B., Schaffner,S.F., Nguyen,H., Moore,J.M., Roy,J., Blumenstiel,B., Higgins,J., DeFelice,M., Lochner,A., Faggart,M. *et al.* (2002) The structure of haplotype blocks in the human genome. *Science*, **296**, 2225–2229.
7. Fredman,D., Siegfried,M., Yuan,Y.P., Bork,P., Lehvaslaiho,H. and Brookes,A.J. (2002) HGVbase: a human sequence variation database emphasizing data quality and a broad spectrum of data sources. *Nucleic Acids Res.*, **30**, 387–391.
8. Bailey,J.A., Gu,Z., Clark,R.A., Reinert,K., Samonte,R.V., Schwartz,S., Adams,M.D., Myers,E.W., Li,P.W. and Eichler,E.E. (2002) Recent segmental duplications in the human genome. *Science*, **297**, 1003–1007.
9. Bedell,J.A., Korf,I. and Gish,W. (2000) MaskerAid: a performance enhancement to RepeatMasker. *Bioinformatics*, **16**, 1040–1041.
10. Chen,L.Y., Lu,S.H., Shih,E.S. and Hwang,M.J. (2002) Single nucleotide polymorphism mapping using genome-wide unique sequences. *Genome Res.*, **12**, 1106–1111.
11. Pearson,W.R. (2000) Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.*, **132**, 185–219.
12. Zdobnov,E.M., Lopez,R., Apweiler,R. and Etzold,T. (2002) The EBI SRS server—new features. *Bioinformatics*, **18**, 1149–1150.
13. Levene,M. and Loizou,G. (2003) Why is the snowflake schema a good data warehouse design? *Inf. Syst.*, **28**, 225–240.
14. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G., Korf,I., Lapp,H. *et al.* (2002) The BioPerl toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.
15. Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
16. Dowell,R.D., Jokerst,R.M., Day,A., Eddy,S.R. and Stein,L. (2001) The Distributed Annotation System. *BMC Bioinformatics*, **2**, 7.
17. Wilkinson,M.D. and Links,M. (2002) BioMOBY: an open source biological web services proposal. *Brief. Bioinform.*, **3**, 331–341.